Advancing Translation Preference Modeling with RLHF: A Step Towards Cost-Effective Solution

Anonymous ACL submission

Abstract

Faithfulness, expressiveness, and elegance is the constant pursuit in machine translation. However, traditional metrics like *BLEU* do not strictly align with human preference of translation quality. In this paper, we explore leveraging reinforcement learning with human feedback (*RLHF*) to improve translation quality. It is non-trivial to collect a large high-quality dataset of human comparisons between translations, especially for low-resource languages. To address this issue, we propose a cost-effective preference learning strategy, optimizing reward models by distinguishing between human and machine translations. In this manner, the reward model learns the deficiencies of machine translation compared to human and guides subsequent improvements in machine translation. Experimental results demonstrate that *RLHF* can effectively enhance translation quality and this improvement benefits other translation directions not trained with *RLHF*. Further analysis indicates that the model's language capabilities play a crucial role in preference learning. A reward model with strong language capabilities can more sensitively learn the subtle differences in translation quality and align better with real human translation preferences.

1 Introduction

As a crucial technology facilitating communication between disparate languages and cultures, machine translation has long garnered significant attention from both academia and industry (Yang et al., 2020). Recently, the emergence of large language models (LLMs) has propelled the field to new frontiers (Yang et al., 2023; Zhu et al., 2023; Jiao et al., 2023b; Hendy et al., 2023). Pre-training on massive monolingual datasets has alleviated the reliance on extensive parallel corpora while enhancing translation quality (Xu et al., 2024).

To enhance the translation capabilities of models, much of the research works have adopted one of two optimization objectives: one is through supervised fine-tuning of translation models to maximize the log probability of human translations (Yang et al., 2023; Xu et al., 2024); the other is through the techniques like reinforcement learning, directly optimizing the similarity score (e.g., *BLEU* score (Papineni et al., 2002)) between model outputs and human translations (Ranzato et al., 2016; Wu et al., 2018; Wieting et al., 2019). Although both approaches have generally performed well, the objectives they optimize for are not fully aligned with human's preferences for translation faithfulness, expressiveness and elegance (Rei et al., 2020; Stiennon et al., 2020).

Fortunately, reinforcement learning from human feedback (RLHF) has been shown to be effective in aligning model behavior with human societal values (Ouyang et al., 2022; Bai et al., 2022). This process integrates reward modeling, where human annotators rank different responses from models based on their preferences, and then normalizes model behavior through a reinforcement learning (RL) phase. However, it is non-trivial to collect a large high-quality preference dataset. Firstly, preference data often comes with noise and ambiguity, as there is low consistency among different human annotators (Wang et al., 2024). More importantly, preference data annotation for translation tasks places higher demands on annotators' linguistic capabilities, a challenge particularly pronounced in low-resource languages.

This paper explores improving translation quality through RLHF and proposes a cost-effective preference learning strategy. We avoid the need to construct expensive preference datasets and instead leverage the inductive bias that *high-quality human translations are superior to machine-generated translations.* The reward model learns human translation preferences by comparing the quality difference between the two, and subsequently guides the improvement of machine translation quality. To collect such high-quality human translations, we align books with multilingual versions. Our motivation for choosing books as the data source is as follows: 1) the original text is authored by writers and the target language is translated by professional translators, ensuring the quality of both texts; 2) compared to web text, book text typically contains more complex language structures, which is particularly beneficial for learning translation preferences; 3) aligning book text does not require as high a level of linguistic capabilities from annotators and can be assisted with external tools (Wang et al., 2023). The experimental results indicate that the reward model effectively learns human translation preferences, and the translation quality of the model is significantly improved.

The main contributions of this paper are as follows: 1) We explore the use of RLHF to improve machine translation quality and propose a cost-effective preference learning strategy that avoids the need for expensive preference data construction; 2) Our experimental results demonstrate that RLHF can improve translation quality, and this improvement can be transferred to target languages not trained with RLHF; 3) Further analysis shows that reward models with strong language capabilities can more sensitively learn differences in translation quality and have stronger resistance to noise in the data.

2 Related works

2.1 Reinforcement Learning from Human Feedback

In recent years, research applying RLHF techniques to tasks involving LLMs has significantly increased (Ouyang et al., 2022; Touvron et al., 2023b), aiming to align the behavior of these models more closely with human preferences. For instance, Stiennon et al. (2020) employ this technique to enhance the quality of summaries, while Bai et al. (2022) utilize it to enable the model to generate responses that are more harmless and useful.

These technique follows a systematic approach: firstly, collect task-specific human preference data. Then, use this data to train a reward model, which acts as a proxy for human preferences. During reinforcement learning, this reward model provides signals to guide model training. However, collecting human preference data is non-trivial, time-consuming, and labor-intensive, often requiring high demands on annotators and plagued by inconsistencies in annotation standards among them. (Bai et al., 2022; Casper et al., 2023; Wang et al., 2024)

2.2 Human-like Alignment in Translation

Achieving human-level machine translation has long been a research goal, receiving ongoing attention. (Hassan et al., 2018; Wu et al., 2016; Läubli et al., 2018) Recent years, some studies have focused on improving the quality of machine translation through human feedback and alignment techniques. Kreutzer et al. (2018) gather implicit task-based feedback, enhancing individual word translations and automatic evaluation measures. Jiao et al. (2023a) employs contrastive instruction and error-guided instruction to align LLMs with human feedback. He et al. (2024) attempt to leverage the quality estimation model as the reward model to predict human preference feedback.

Considering the methods above, the scarcity of human-preference data in translation has long been a bottleneck. Our approach differs, creatively utilizing meticulously translated human data as readily available preference data.

3 Improving Translation with RLHF

To build a translation model that aligns with human translation preferences, we start with a generic pre-trained language model π^{pre} (such as LLaMA (Touvron et al., 2023a)), and follow the pipeline of the following three steps: 1) Supervised fine-tuning of π^{pre} on parallel corpora yields a model π^{sft} with basic translation capabilities; 2) Training a reward model r on preference dataset $\mathcal{D}_{\text{rm}}$, which assigns high reward scores to translations that adhere to human preference; 3) Utilizing r as a proxy for human preferences, enhancing the translation quality of the model through reinforcement learning.

3.1 Supervised Fine-tuning to Acquire Basic Translation Capabilities

Given a parallel corpus $\mathcal{D}_{\text{sft}} = \{(x^{(i)}, y^{(i)})\}_{i=1,..,n}$, where x_i represents the source-language text and y_i represents the corresponding reference translation, we utilize a fixed prompt template $\mathcal{I}$ and construct the training data as follows:

$\mathcal{I}$ =“Translate this from [SRC] to [TGT]:
[SRC]: <x> [TGT]: <y>”

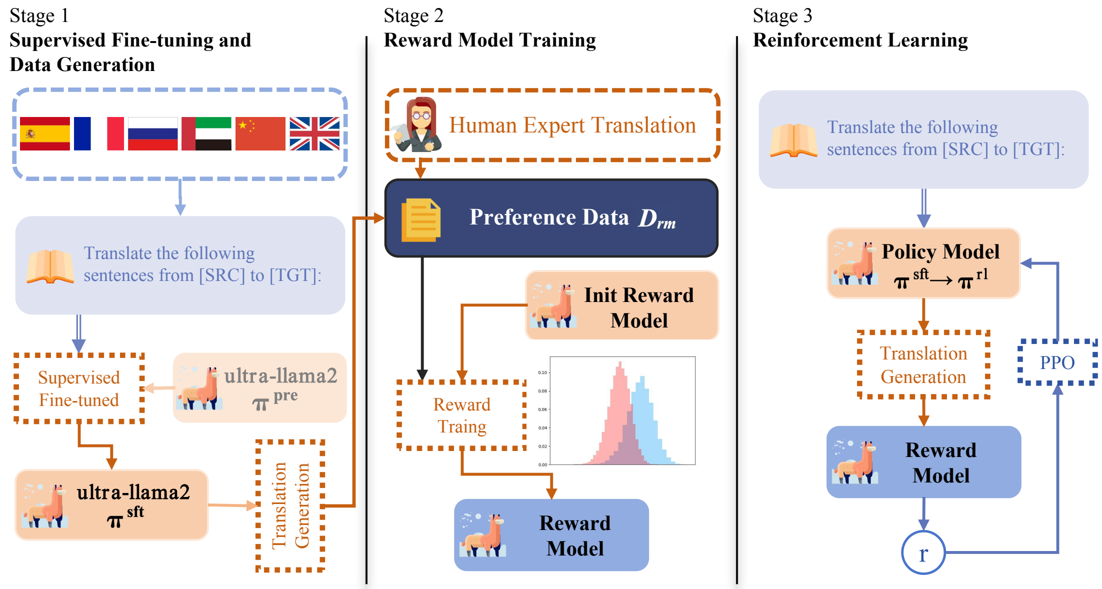

Figure 1: An Overview of Modeling Translation Preferences using RLHF; To achieve cost-effective preference learning, we optimize the reward model in the second step by contrasting the deficiencies of SFT model translations with human expert translations, thus avoiding the expensive labeling of preference data.

where, 'SRC' and 'TGT' respectively represent the names of the source language and the target language. The translation model π^{sft} is optimized via the negative log-likelihood loss on parallel corpus $\mathcal{D}^{\text{sft}}$ as follows:

$$\mathcal{L}_{NLL} = -\mathbb{E}_{(x,y)\sim\mathcal{D}^{\text{sft}}} \log \pi^{\text{sft}}(y|x,\mathcal{I}), \quad (1)$$

The translation model π^{sft} acquired basic translation capabilities by maximizing the probability of reference translations.

3.2 Modeling Translation Preferences

To accurately model human preferences, high-quality preference data is crucial. A common practice used for modeling human value preferences is to prompt the model to generate two different outputs $(y1, y2) \sim \pi^{\text{sft}}(\cdot|x)$ in response to a query x and then require annotators to choose their preferred one, i.e., $y_w > y_l$. y_w and y_l denote the chosen and rejected response, respectively. However, constructing a large preference dataset for translation tasks requires annotators who are experts/native speaker in the specific languages, which greatly increases the annotation cost. For low-resource languages, finding a sufficient number of qualified annotators may even be impractical.

Unlike the aforementioned approach, we instead leverage the induction bias of '*high-quality human translation is superior to machine-generated translation*' to collect preference data at a lower cost. These high-quality human translations are sourced from book data. Our motivation for selecting this data source is as follows: 1) Books' original texts and their translated versions are completed by authors and professional translators, ensuring high text quality; 2) Book corpora contain more complex language structures compared to web text, which is highly beneficial for preference learning; 3) Aligning book data requires less stringent language proficiency from annotators and can be aided by external tools.

We optimize our reward model r by contrasting the differences between high-quality human translation and machine translation:

$$\mathcal{L}(r) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}_{\text{rm}}}[log\sigma(r(x,y_w)-r(x,y_l))], \quad (2)$$

where x represents the source language sentence, while y_w and y_l respectively denote a high-quality human translation and a machine-generated translation, and $\mathcal{D}_{\text{rm}} = \{(x^{(i)}, y_w^{(i)}, y_l^{(i)})\}_{i=1,..,N}$ is the preference dataset.

3.3 Improving Translation via RL Fine-tuning

During the Reinforcement Learning (RL) phase, we employ the acquired reward function to furnish feedback to the language model. Specifically, we

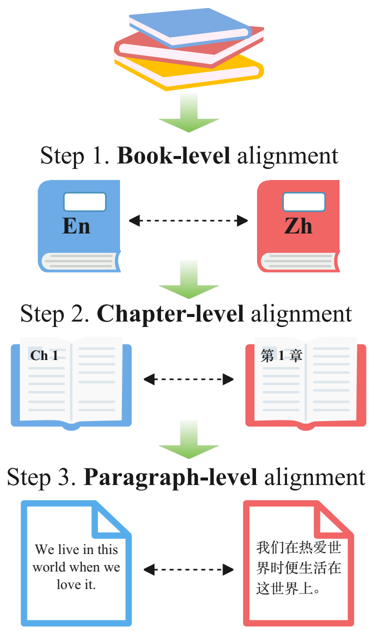

Figure 2: The process of constructing the English-Chinese book dataset.

refine the policy model to optimize the following reward objective:

$$r_{total} = r(x, y) - \eta KL(\pi^{\mathrm{RL}}(y|x)||\pi^{\mathrm{SFT}}(y|x)), \quad (3)$$

where η represents a coefficient regulating the extent of the KL penalty. The KL divergence component serves two main purposes within this framework. Firstly, it functions as an entropy bonus, maintaining diversity in generation and averting the collapse into singular high-reward responses (Jaques et al., 2019). Secondly, it ensures that the output of the RL policy remains within a distribution where the reward model accurately reflects the performance, thereby preventing significant deviations.

4 Experimental Setup

4.1 Training Data Collection

We collect and utilize translation training data from three different sources. The detailed information of these datasets can be found in table 1.

English-Chinese Books. In order to collect rich human expression habits in book translation data, we manually construct an English-Chinese parallel book corpus dataset. The construction process of this dataset, as shown in Figure 2, can be divided into three steps: Firstly, alignment at the **book level**. We manually collect Chinese and English versions of several books, ensuring high quality for both versions selected, with translations being provided by skilled professional translators. Next, alignment at the **chapter level** is performed for each book's Chinese and English versions. We parse the data of the entire book into text format and then compare the number and content of chapters for consistency. Finally, we align Chinese and English paragraphs at the **paragraph level** for each chapter through manual comparison and adjustment.

Yiyan Corpus.[1] To enhance the diversity of the data and strengthen the model's robustness to inputs of different lengths, we incorporate the Yiyan corpus, an English-Chinese Parallel Corpus. Specifically, we utilize the academic and novel sections, consisting of parallel sentences translated by human translators at the sentence level.

United Nations Parallel Corpus (UN). (Ziemski et al., 2016) For our multilingual experiments, we use the UN training set, which was also manually translated. This dataset includes parallel data in six languages: English, Chinese, French, Spanish, Russian, and Arabic. We conduct experiments on translation from English to the other five languages. We randomly sample from the extensive dataset, ensuring English sentences contain a minimum of 30 words to guarantee richer information.

In the experiment for bidirectional English-Chinese translation, we mix English-Chinese books data with Yiyan Corpus data. For the multilingual experiment, we utilize the UN dataset.

4.2 Model

- Ultra-LLaMA2-7B: Base model of our experiments. A variant of LLaMA2-7B further-pretrained on over $200B$ Chinese tokens.
- LLaMA2-7B (Touvron et al., 2023b): A LLM trained primarily in English. In certain experiment, we use this model as the control.

4.3 Evaluation

4.3.1 Metrics

When evaluating the quality of translation results, we employed three evaluation methods: GPT-4 comparative evaluation (OpenAI, 2023) and

[1] https://corpus.bfsu.edu.cn/info/1070/1631.htm

Name of the dataset	Translation direction	Granularity	Training Samples
English-Chinese Books	En $\Leftrightarrow$ Zh	paragraph-level	$60,000$
Yiyan Corpus	En $\Leftrightarrow$ Zh	sentence-level	$30,000$
United Nations Parallel Corpus	En $\Rightarrow$ Zh/ Fr/ Es/ Ru/ Ar	sentence-level	$60,000$

Table 1: Details of translation training data. In English-Chinese Books dataset and Yiyan Corpus dataset, we simultaneously use both directions of parallel corpora. In United Nations Parallel Corpus, we utilize approximately $60,000$ samples from English to each language.

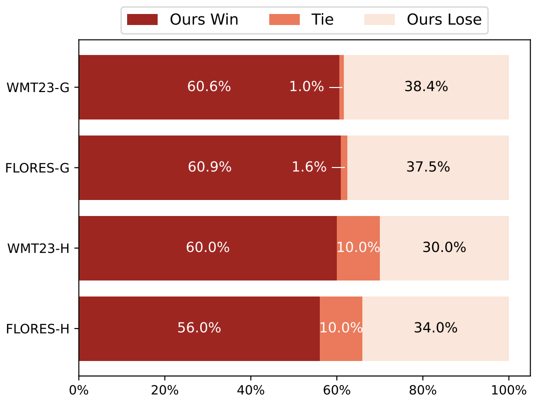

Figure 3: Comparison between preference optimized models and the SFT model on Task En→Zh. G and H represent GPT-4 and humans as evaluators, respectively.

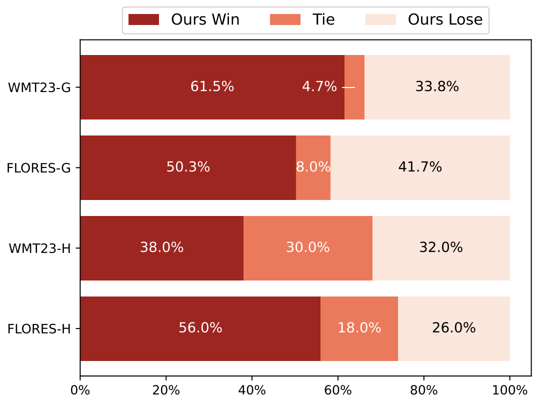

Figure 4: Comparison between preference optimized models and the SFT model on Task Zh→En. G and H represent GPT-4 and humans as evaluators, respectively.

COMET metrics (Rei et al., 2020) and human evaluation.

GPT-4. Due to its exceptional general-purpose capabilities, the GPT-4 model has emerged as a pioneering approach for evaluating NLP tasks. We present the original text of a given sentence alongside translations from both the SFT and RLHF models, allowing GPT-4 to compare them simultaneously and select the superior translation. In the prompt used during the tests, we explicitly included multidimensional evaluation criteria, including flexibility, fidelity, and accuracy and so on. To mitigate the impact of comparison order, we interchanged the positions of both models' outputs for each test, conducting two evaluations simultaneously. Refer to the Table 5 in appendix for the complete prompt.

COMET. COMET is a neural framework for training multilingual machine translation evaluation models. It has been shown to have high correlation with human assessment and has become an increasingly widely used metric for machine translation evaluation (Kocmi et al., 2021). We select the reference-free quality evaluation model wmt22-cometkiwi-da Rei et al. (2022). We compare the translation abilities of two models (SFT and RLHF models) by evaluating the relative COMET scores of their translation results for the same translated data.

Human Evaluation. When evaluating bidirectional English-Chinese translation, we also incorporate human evaluation. Proficient bilingual native speakers conduct assessments to compare translation quality.

4.3.2 Test Sets

We utilize the WMT23 test sets (Kocmi et al., 2023) and the Flores-200 devtest sets (Costa-jussà et al., 2022) to assess the model's performance. Note that WMT23 does not cover all directions for the multilingual experiment, but as we employ comparative reference-free evaluation, we only use English data from the WMT23 test sets as the source.

5 Results and Disscussions

5.1 Main Results

Is it feasible to model translation preferences without explicit preference annotations?
This paper explores the feasibility of modeling

Faithfulness	Input	The synthesis of the pharmaceutical compound acetylsalicylic acid, commonly known as aspirin, marked a significant advancement in modern medicine.
	SFT	阿司匹林的合成标志着现代医学的一个重要进步。
	RLHF	乙酰水杨酸（阿司匹林）这种药物的合成，标志着现代医学的一个重要进步。
	Commentary	In the translation by RLHF, the term ‘乙酰水杨酸这种药物’ corresponds to ‘the pharmaceutical compound acetylsalicylic acid’ in the input text, while in the translation by SFT, this expression is missing, reflecting an improvement in translation faithfulness.
Expressiveness	Input	After years of practice, running a marathon was a piece of cake for her.
	SFT	经过多年的练习，对她来说，跑马拉松就像吃蛋糕一样简单。
	RLHF	经过多年的锻炼，跑马拉松对她来说已是小菜一碟了。
	Commentary	In the SFT translation, ‘像吃蛋糕一样简单’ is a literal translation of "a piece of cake" in the input text. In contrast, the translation in RLHF, ‘小菜一碟’, is a more authentic Chinese expression, vivid and expressive. This case reflecting an enhancement in the expressive power of the translation.
Elegance	Input	As the crimson hues of dusk melded with the cerulean tapestry of the night sky, the poet pondered over verses that could encapsulate the ephemeral beauty of the twilight.
	SFT	夜幕降临，天空中的蓝色帷幕与黄昏的红色调和在一起，诗人开始思考如何用诗句来捕捉这短暂的美好。
	RLHF	暮色渐浓，绯红的余晖与夜空的青蓝交织，诗人思忖着如何用诗句来捕捉这转瞬即逝的美景。
	Commentary	Both ‘转瞬即逝’ and ‘短暂’ can be used to convey the meaning of ‘ephemeral’ in the input text, but the former implies a sense of regret and sorrow for the fleeting nature of beautiful things, while the latter is a neutral term, simply describing temporal brevity. This example demonstrates an improvement in the elegance of the translation.

Table 2: An case study on modeling human translation preference through RLHF. The yellow background text reflects the improved translation quality of RLHF compared to SFT.

human translation preferences in the absence of explicit preference annotations. By comparing the deficiencies of machine translation with human translation, the reward model learns human translation preferences, thus circumventing the need for costly preference data annotation. In this subsection, we empirically validate the effectiveness of this approach. Specifically, we use high-quality English-Chinese parallel corpora (refer to Section 4.1) as preferred data, while data generated by the SFT model (also fine-tuned using pre-heldout book data) serves as dispreferred data. From Figure 3 and 4, we observe that on the WMT23 and FLORES datasets, our preference-optimized model exhibits significantly improved win rates compared to the SFT model, regardless of whether the evaluator is GPT-4 or human. This indicates that with access to high-quality parallel corpora, even in the absence of explicit preference annotations, we can learn human translation preferences and improve the translation quality of the model. In Table 2, we demonstrate the quality improvement of translations after preference optimization through three cases.

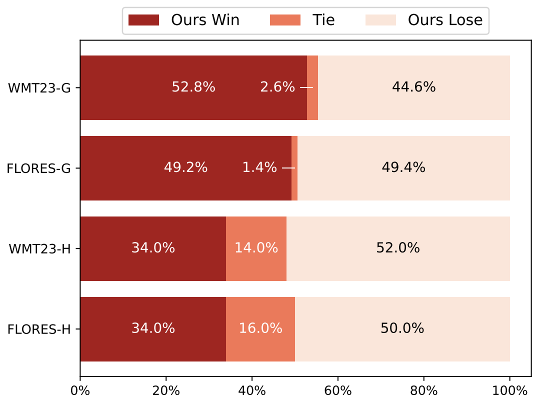

Figure 5: After replacing the base model in Figure 3 with LLaMA, compare the preference optimized model and the SFT model in the En→Zh translation direction.

The language capability of reward model is crucial for preference learning. In the previous part of the experiment, we utilize Ultra-LLaMA as the base model, which is a variant of LLaMA further-pretrained on over $200B$ Chinese tokens. To investigate the impact of language capability differences on preference learning, we replace the base model with original LLaMA, which has a relatively weaker processing capability

Dataset	Evaluator	Results	Translation Direction				
			En→Fr	En→Es	En→Ru	En→Zh	En→Ar
WMT23	GPT-4	SFT Win	**0.510**	0.432	0.462	0.395	0.447
		RLHF Win	0.430	**0.439**	**0.490**	**0.552**	**0.534**
		Tie	0.060	0.129	0.048	0.053	0.019
	COMET	SFT Win	0.416	0.386	0.450	0.326	0.450
		RLHF Win	**0.544**	**0.506**	**0.516**	**0.634**	**0.550**
		Tie	0.040	0.108	0.034	0.040	0.000
FLORES	GPT-4	SFT Win	**0.495**	0.378	0.455	0.347	0.416
		RLHF Win	0.417	**0.396**	**0.477**	**0.587**	**0.552**
		Tie	0.088	0.226	0.068	0.066	0.032
	COMET	SFT Win	0.398	0.344	0.424	0.328	0.448
		RLHF Win	**0.536**	**0.472**	**0.526**	**0.624**	**0.552**
		Tie	0.066	0.184	0.050	0.048	0.000

Table 3: Results of preference modeling in five translation directions on the UN dataset.

for Chinese. We construct the SFT model using the same experimental data and training scheme as in the previous section and further optimize it for human preferences. As observed from Figure 5, the win rate of the preference-optimized model significantly decreased in comparison with the SFT model, and it even lost to the SFT model in human evaluations. It is worth noting that the SFT model trained on original LLaMA inherently lacks translation capabilities compared to the SFT model based on Ultra-LLaMA, thus highlighting more pronounced differences in the quality of generated translations compared to human translations. Intuitively, this should decrease the learning difficulty of the reward model. However, the reward model constructed based on original LLaMA failed to effectively model human translation preferences. Therefore, we believe that the language capability of reward models plays an important role in preference learning.

5.2 The Impact of the Inherent Nature of Human Translation

The book dataset used in the previous section has high textual quality, containing complex linguistic structures and grammar phenomena, and is diverse in its domain sources. In contrast, the UN originates from specific domains and lacks complex linguistic structures and rhetorical devices commonly found in governmental documents. In this section, we conduct multilingual experiments using the UN dataset to explore the influence of intrinsic properties of the data on preference learning.

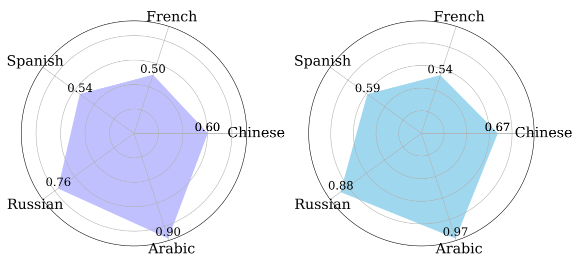

Figure 6: Quality Analysis of UN Datasets.

For simple domain-specific parallel corpora, the quality of machine translations is comparable to human translations.

As shown in Figure 6 (left), using COMET as the evaluation metric, we find that the difference in quality between translations from the SFT model and human translations is minimal. Especially for French and Spanish, only 50% and 54% of human translations respectively outperform translations from the SFT model. This indicates that when parallel corpora do not contain complex linguistic sources or sentence structures, the SFT model can already achieve results comparable to human translations. Clearly, the induction bias of "human translations are superior to translations from the SFT model" is no longer valid for such datasets.

Similar translation quality increases the difficulty of preference learning.

To explore preference learning on the United Nations dataset, we first remove 50% of the data with small differences in COMET scores,

Translation Direction Optimized by RLHF	Evaluator	Results	Transferred Translation Direction				
			En→Fr	En→Es	En→Ru	En→Zh	En→Ar
En→Zh	GPT-4	SFT Win	0.443	0.448	0.418	–	0.355
		RLHF Win	**0.540**	**0.493**	**0.563**	–	**0.563**
		Tie	0.018	0.030	0.020	–	0.083
	COMET	SFT Win	0.390	0.410	0.475	–	0.420
		RLHF Win	**0.610**	**0.590**	**0.525**	–	**0.580**
		Tie	0.000	0.000	0.000	–	0.000
En→Ar	GPT-4	SFT Win	0.458	**0.465**	0.455	**0.485**	–
		RLHF Win	**0.510**	0.458	**0.533**	**0.485**	–
		Tie	0.033	0.078	0.013	0.030	–
	COMET	SFT Win	0.410	**0.505**	0.435	**0.580**	–
		RLHF Win	**0.590**	0.495	**0.565**	0.420	–
		Tie	0.000	0.000	0.000	0.000	–

Table 4: Cross-lingual Transfer Results of Translation Preferences.

retaining data pairs with relatively clear preference tendencies. However, as shown in Figure 6 (right), in the directions of French and Spanish, nearly 50% of SFT translations still outperform human translations. Therefore, we reannotate based on COMET scores to construct a preference dataset. As shown in Table 3, translation models optimized for preferences significantly outperform the SFT model in all five translation directions in terms of COMET scores. This is easily understood since our preference labels are derived from COMET scores. However, learned preferences may not necessarily be generalizable and aligned with human preferences. The evaluation results of GPT-4 in Table 3 indicate that in the English to Spanish and Russian directions, the preference-optimized model only has a slight advantage, and in the case of French, it even loses to the SFT model. This is mainly because the difference between SFT and human translations is minimal in French. In contrast, in the English to Arabic direction, the preference-optimized model consistently and significantly improves, mainly due to the distinct differences in preference data itself, making it easier for the reward model to learn generalizable translation preferences.

5.3 Transferability Analysis

With the powerful Chinese capabilities of the reward model and the notable quality disparities in Arabic preference data, translation models have achieved effective alignment with human preferences in both English-to-Chinese and English-to-Arabic directions. In this section, we explore through experiments whether learned translation preferences can be transferred across languages. As observed from Table 4, RLHF training solely on tasks in English-to-Chinese translation, the learned human preferences can effectively transfer to other languages and consistently improve performance. Similarly, when English-to-Arabic translation is used as the source task, improvements are also evident in tasks such as English-to-French and English-to-Russian translation. This indicates that aligning with and transferring from human preferences in other translation directions can be a viable strategy when the current translation direction lacks reward models with strong language capabilities or high-quality preference data.

6 Conclusions

This paper explores modeling translation preferences with RLHF to improve the quality of machine translation. We propose a cost-effective preference learning strategy, optimizing reward models by contrasting deficiencies in machine translation compared to human translation. Learning human preferences while avoiding expensive preference data annotation. Further analysis suggests that the language capability of the reward model and the nature of the data itself affect the effectiveness of preference learning. Additionally, learned preferences exhibit cross-lingual transfer phenomena. This may be beneficial for preference modeling in low-resource languages.

Limitations

Due to cost limitations, we only collected English-Chinese aligned book data as a substitute for preference data, without covering more translation directions. Additionally, our human evaluations were limited to English-Chinese translation, with GPT-4 used as a proxy for manual evaluations in other translation directions. In the future, we will attempt to align with human translation preferences in more languages, especially low-resource languages, and conduct comprehensive manual evaluations in more translation directions.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Ségerie, Micah Carroll, Andi Peng, Phillip J. K. Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Biyik, Anca D. Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *CoRR*, abs/2307.15217.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. *CoRR*, abs/1803.05567.

Zhiwei He, Xing Wang, Wenxiang Jiao, Zhuosheng Zhang, Rui Wang, Shuming Shi, and Zhaopeng Tu. 2024. Improving machine translation with human feedback: An exploration of quality estimation as a reward model. *CoRR*, abs/2401.12873.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation.

Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Àgata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind W. Picard. 2019. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *CoRR*, abs/1907.00456.

Wenxiang Jiao, Jen-tse Huang, Wenxuan Wang, Zhiwei He, Tian Liang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023a. Parrot: Translating during chat using large language models tuned with human translation and feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 15009–15020. Association for Computational Linguistics.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023b. Is chatgpt a good translator? yes with gpt-4 as the engine.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondrej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popovic, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): llms are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation, WMT 2023, Singapore, December 6-7, 2023*, pages 1–42. Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation, WMT@EMNLP 2021, Online Event, November 10-11, 2021*, pages 478–494. Association for Computational Linguistics.

Julia Kreutzer, Shahram Khadivi, Evgeny Matusov, and Stefan Riezler. 2018. Can neural machine translation be improved with user feedback? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 3 (Industry Papers)*, pages 92–105. Association for Computational Linguistics.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? A case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4791–4796. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2685–2702. Association for Computational Linguistics.

Ricardo Rei, Marcos V. Treviso, Nuno Miguel Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte M. Alves, Luísa Coheur, Alon Lavie, and André F. T. Martins. 2022. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation, WMT 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 634–645. Association for Computational Linguistics.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. 2020. Learning to summarize from human feedback. *CoRR*, abs/2009.01325.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Binghai Wang, Rui Zheng, Lu Chen, Yan Liu, Shihan Dou, Caishuang Huang, Wei Shen, Senjie Jin, Enyu Zhou, Chenyu Shi, Songyang Gao, Nuo Xu, Yuhao Zhou, Xiaoran Fan, Zhiheng Xi, Jun Zhao, Xiao Wang, Tao Ji, Hang Yan, Lixing Shen, Zhan Chen, Tao Gui, Qi Zhang, Xipeng Qiu, Xuanjing Huang, Zuxuan Wu, and Yu-Gang Jiang. 2024. Secrets of rlhf in large language models part ii: Reward modeling.

Longyue Wang, Zefeng Du, DongHuai Liu, Deng Cai, Dian Yu, Haiyun Jiang, Yan Wang, Shuming Shi, and Zhaopeng Tu. 2023. Guofeng: A discourse-aware evaluation benchmark for language understanding, translation and generation.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond BLEU:training neural machine translation with semantic similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355, Florence, Italy. Association for Computational Linguistics.

Lijun Wu, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2018. A study of reinforcement learning for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3612–3621, Brussels, Belgium. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim

Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models.

Shuoheng Yang, Yuxin Wang, and Xiaowen Chu. 2020. A survey of deep learning techniques for neural machine translation.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. Bigtranslate: Augmenting large language models with multilingual translation capability over 100 languages.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual machine translation with large language models: Empirical results and analysis.

Michal Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

A Implementation Details

SFT stage. In the English-Chinese model, we use $1/3$ of the dataset, with a learning rate of $5e-6$, training for 2 epochs; In the multilingual model, approximately $3/4$ of the training data is used for 1 epoch, with a learning rate of $5e-6$.

RM training stage. The reward model is initialized with the previous stage's SFT model. In the English-Chinese model, the remaining $2/3$ of the training data are used to form chosen-rejected pairs with the data generated by the SFT model; In the multilingual model, the remaining $1/4$ of the training data is utilized, and only the top 50% of high-confidence data selected by the COMET model, is used to train the RM. Training continues with dynamic batch processing until early stopping criteria are met.

RL stage. For English-Chinese model, we reuse the inputs from the RM stage's training data as queries, and for multilingual model, we use English monolingual book data obtained from web crawling as queries. We set the KL divergence penalty coefficient to 0.02, and trained until early stopping criteria were met.

You are a translation expert, and I need your help in impartially judging the quality of two translations. The judging criteria are as follows: Flexibility of Translation: A good translation is not confined to the original form, and it should be smooth and clear. Poor-quality translations appear rigid and awkward, merely translating word-for-word according to the original form. Fidelity of Translation: A good translation should faithfully reflect the content of the original text. It should not introduce content that does not exist in the original, nor should it omit content present in the original. Accuracy and Elegance of Phrasing: In a good translation, phrases and wording should adhere to the conventions of the target language, and they should be as accurate and elegant as possible. Next, I will provide you with the original text and two translations. Please let me know which one is better according to these criteria. Please give your judgment directly and do not output additional explanations.

Table 5: Prompt template for GPT4 evaluaiton.